Universidade de Brasília
Departamento de Estatística

Estudo das Zonas Especiais de Interesse Social através da metodologia de análise de sobrevivência

Paloma Emanoelle de Araujo Ribeiro

Relatório Final de Monografia apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Paloma Emanoelle de Araujo Ribeiro

Estudo das Zonas Especiais de Interesse Social através da metodologia de análise de sobrevivência

Orientadora:
Profa. Dra. **Juliana Betini Fachini Gomes**

Relatório Final de Monografia apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Agradecimentos

Agradeço à Deus por sua graça, bondade e pelo amor infinito concedidos, que me guiaram por toda a vida e que se não fosse pela vontade D'Ele, nada disso teria sido realizado.

À minha família pela torcida e pelo apoio, aos meus irmãos pela compreensão e por ser uma das bases mais importantes da minha existência, aos meus pais e especialmente a minha mãe Irani, por ser o meu alicerce, pela força transmitida, por lutar pela minha felicidade, por estar presente em todos os momentos da minha vida e finalmente, por ser o meu maior exemplo de integridade, caráter e competência. Saiba que quando eu crescer, quero ser igual a você.

Ao meu namorado, Paulo, pelo amor, por segurar a minha mão em momentos difíceis, por sempre acreditar no meu potencial e pelo companheirismo que nos mantém unidos desde 2012 e a sua família, que também é minha, por todo carinho dedicado.

À Jady por tornar o desenvolvimento deste trabalho um fardo mais leve com a sua amizade nos últimos semestres e aos meu amigos queridos que sempre se alegraram com as minhas conquistas e me incentivaram quando era preciso.

Por fim e não menos importante, à professora Juliana Betini Fachini Gomes pela compreensão com a minha rotina, por todo o período de orientação do TCC, pelo apoio, pela preocupação e pela disposição. Como havia dito, desde o primeiro dia de aula na Estatística, eu soube que você seria minha orientadora, e tenho certeza de que foi Deus quem colocou você em meu caminho. Obrigada!

Sumário

Resumo

Estudo das Zonas Especiais de Interesse Social através da metodologia de análise de sobrevivência

Neste trabalho é proposto um modelo de regressão Log-Logístico discreto aplicado a dados de sobrevivência. A motivação se deve à escassez de modelos probabílisticos discretos na área análise de sobrevivência e a especificidade do tipo de variável resposta, que impossibilita, em alguns casos, a aplicação de distribuições contínuas ou a adaptação das mesmas. Os modelos propostos foram aplicados em dois conjuntos de dados referentes à política Zona Especial de Interesse Social, nos quais se verifica a influência de covariáveis no tempo de adesão dos municípios brasileiros à política. Como resultado, foram obtidos três modelos adequados, nos quais dois se referem à adoção da política independente da forma que foi adotada e o terceiro corresponde ao estudo da adesão dos municípios à política apenas por meio de uma legislação específica. Após, foi utilizada uma análise de resíduos com o intuito de verificar a qualidade do ajuste. Todas as análises foram realizadas no *software* R.

Palavras-chave: Dados censurados; Distribuição Log-Logística Discreta; Modelos de regressão.

Abstract

Study of Social Interest Special Zones through the metodology of survival analysis

This work proposes a discrete log-logistic regression model applied to survival data. The motivation is due to the scarcity of probabilistic models in the field of political science in Brazil and the specificity of the type of response variable, which in some cases makes it impossible to apply continuous distributions or adapt them. The proposed models were applied in two datasets referring to the Zona Especial de Interesse Social, in which the influence of covariables on the time of adhesion of Brazilian municipalities to politics is verified. As a result, three suitable models were obtained, in which two refer to the adoption of the independent policy in the form that was adopted and the third corresponds to the study of the adhesion of the municipalities to the policy only by means of a specific legislation. Afterwards, a residue analysis was used in order to verify the quality of the adjustment. All analyzes were made on *software* R.

Keywords: Censored Data; Log-Logistic Discret Distribution; Regression models.

Lista de Figuras

1 Introdução

A estatística consiste no planejamento da pesquisa, coleta, análise e interpretação de dados através de técnicas específicas, com o intuito de propor modelos que expliquem a ocorrência de eventos de interesse e, através desses, auxiliar na tomada de decisão.

A análise de sobrevivência é uma área da estatística que possui uma vasta abrangência de aplicações e está em constante crescimento. Ela consiste no estudo do tempo até a ocorrência de um evento de interesse, também chamada de falha. O grande diferencial entre as técnicas estatísticas convencionais e a análise de sobrevivência é a incorporação de informações incompletas, também chamada de censura.

No trabalho proposto, serão analisados dois conjuntos de dados referentes aos municípios brasileiros: o primeiro corresponde ao tempo em que a política Zona Especial de Interesse Social leva para ser aderida pelos municípios via legislação específica e o segundo corresponde ao tempo que a política leva para ser implantada nos municípios brasileiros, independente da forma adotada (via leis específicas ou Plano Diretor). Para tanto, é necessário definir dois conceitos importantes em relação ao trabalho a ser desenvolvido: Plano diretor e Zona Especial de Interesse Social.

Plano Diretor é o instrumento básico da política de desenvolvimento de um município, instituído pela Constituição Federal de 1988 e possui como finalidade básica "instruir" tanto o poder público quanto a iniciativa privada em relação ao desenvolvimento e crescimento urbano e rural, visando assegurar melhores condições de vida para a população.

As Zonas Especiais de Interesse Social (ZEIS) surgiram a partir da década de 1980 e são definidas como a parcela de área urbana instituída pelo Plano Diretor ou definida por outra lei municipal, destinada preponderantemente à população de baixa renda através de:

- urbanização de bairros ou imóveis públicos;
- aprovação de loteamentos ou desmembramentos;
- regularização de núcleos urbanos informais consolidados.

Assim, o objetivo principal deste trabalho é utilizar as técnicas estatísticas para avaliar o tempo que um município brasileiro leva para a adoção da ZEIS. Esse objetivo será alcançado através de análises descritivas, da busca de um modelo probabilístico que se ajuste adequadamente para os dados avaliados e, posteriormente, inserir covariáveis que possam explicar a adesão dessa política por um município.

O conteúdo abordado neste trabalho está dividido da seguinte forma: Inicialmente tem-se a revisão literária, que compreende toda a teoria utilizada para este estudo. Após, serão apresentados os métodos utilizados para análise na seção de metodologia e, seguindo a estrutura tradicional, serão expostos os resultados e considerações finais. Para a realização de todas as análises, será utilizado o *software* R.

2 Revisão de Literatura

2.1 Análise de Sobrevivência

A Análise de Sobrevivência é uma área da estatística em constante crescimento, pois abrange grande quantidade de aplicações, que vão desde estudos biomédicos a estudos na área de engenharia, seguros e finanças (COLOSIMO;Giolo, 2006).

Na análise de Sobrevivência, a variável resposta é o tempo de falha. Essa variável é definida como o tempo em que a observação levou até atingir um evento de interesse determinado, ou seja, corresponde à diferença entre o tempo em que a falha ocorreu e o tempo inicial do estudo. Essa falha pode representar a morte de um paciente, a cura de determinada doença, a falha de algum equipamento, ou, nesse caso, a adesão de um município à uma política pública. Dependendo da forma que os dados foram coletados, o tempo de falha será descrito por valores discretos, como por exemplo, em anos completos, ou por valores contínuos, como por exemplo, a hora e o dia em que ocorreu a falha. É importante ressaltar que esse tempo é estritamente positivo, uma vez que as falhas passam a ser consideradas somente a partir do tempo inicial.

A principal diferença entre a análise estatística tradicional e a análise de sobrevivência consiste no uso de dados censurados.

Censura corresponde à observações incompletas ou coletadas parcialmente. Essas observações são importantes e devem compor a análise estatística, pois mesmo que estejam incompletos, os dados censurados fornecem informações sobre o tempo de vida do elemento. A retirada desses dados acarretaria em resultados viciados. Dito isto, é necessário introduzir no estudo uma variável indicadora de censura, representada por $\delta_i,\ i = 1, 2, 3, ..., n$:

$$\delta_i = \begin{cases} 0 & \text{, se a } i\text{−ésima unidade de informação foi censurada} \\ 1 & \text{, se a } i\text{−ésima unidade de informação falhou.} \end{cases}$$

Há três tipos de censura a serem considerados:

- Censura tipo I: Ocorre em estudos com tempos de coleta pré-estabelecidos, quando os elementos não alcançam o evento de interesse até o fim do estudo.

- Censura tipo II: Em alguns casos, é interessante para o pesquisador, que ocorra um determinado número de falhas antes do fim do estudo. Assim, todos os elementos que, tendo atingido o evento de interesse ou não, ultrapassarem esse número de falha, são censurados, tornando, nesse caso, o percentual de censuras um valor constante.

- Censura Aleatória: A censura aleatória ocorre quando um elemento é retirado do estudo sem ter ocorrido a falha, ou se falhou por causa diferente da qual o estudo está interessado, ou seja, o pesquisador não tem controle do número de censuras no estudo.

Os tipos de censura aqui apresentados são conhecidos por censura à direita, ou seja, o tempo de ocorrência do evento de interesse é sempre maior que o tempo observado. Há também outras formas de censura: censura à esquerda e censura intervalar.

Censura à esquerda ocorre quando o tempo observado é maior que o tempo de falha do elemento. Ou seja, o evento de interesse ocorreu antes de o elemento ser observado. Censura intervalar é um tipo mais geral, onde o evento de interesse ocorre em um intervalo de tempo.

Neste trabalho o foco estará na censura à direita do tipo I, pois devido ao comportamento dos dados, o município só pode ser censurado se o estudo acabar, tornando assim, todos os outros tipos inviáveis.

2.2 Caso 1: Variáveis aleatórias discretas

Em análise de sobrevivência, dependendo da estrutura dos dados, o tempo de ocorrência de falha apresenta valores inteiros e não-negativos, ou seja, a variável de interesse corresponde a uma variável de contagem, com tempos $t = \{0, 1, 2, 3, ...\}$. Nesse caso, as principais funções utilizadas na área de análise de sobrevivência serão definidas nas próximas subseções.

2.2.1 Função de distribuição de probabilidades

É importante definir a função de distribuição de probabilidades da variável aleatória T. Seja o T tempo de ocorrência de falha de uma observação uma variável aleatória não-negativa, a função $p(t) = P(T = t)$ deve satisfazer as seguintes condições:

- $p(t_i) \geq 0$,
- $\sum_{t=0}^{\infty} p(t) = 1$.

2.2.2 Função de sobrevivência

A função de sobrevivência é a base da análise estatística do tempo de falha, pois dá suporte a todas as outras funções, que serão utilizadas posteriormente. Ela representa a

probabilidade de um determinado elemento do estudo não falhar até um tempo t e é dada por (Nakano,2017):

$$S(t) = P[T > t] = \sum_{k=t+1}^{\infty} p(k) = \sum_{k=t+1}^{\infty} P(T = k), t = 0, 1, 2, ...$$

Vale lembrar que a função de sobrevivência é definida para todos os números reais não-negativos, isto é, $S(t)$ decresce para todos os pontos onde t tem probabilidade positiva e é constante nos demais pontos (ou seja, para variáveis aleatórias discretas, $S(t)$ assume uma forma de "escada").

2.2.3 Função de risco (Taxa de falha)

No caso de variável aleatória discreta, a função de risco (ou taxa de falha) corresponde à probabilidade condicional do elemento experimentar o evento de interesse num instante t dado que sobreviveu (ou seja, não ocorreu falha) até t, descrita como:

$$h(t) = P(T = t \mid T \geq t), t = 0, 1, 2, ...$$

É importante ressaltar que, para valores de t não-inteiros ou negativos, a função de risco é zero. Porém, ela pode assumir diversas formas e é mais informativa que a função de sobrevivência, pois a forma da função de risco pode, a priori, trazer um palpite inicial de qual função de distribuição o modelo irá seguir, além de ser mais flexível do que a função de sobrevivência, uma vez que não é limitada superiormente.

2.2.4 Função de risco acumulado

A função de risco acumulado não possui interpretação direta, mas se torna útil nos casos em que a estimação de $h(t)$ é difícil, sendo definida por:

$$H(t) = \sum_{k=0}^{t} h(k).$$

2.3 Caso 2: Variáveis aleatórias contínuas

Quando o tempo de falha é medido de forma "contínua", como por exemplo, em horas e minutos, a variável aleatória T passa a assumir valores contínuos e não-negativos.

2.3.1 Função densidade

Seja, neste caso, a variável resposta tempo de falha, denotada por T, uma variável aleatória contínua, estritamente positiva. Essa variável possui uma função densidade de probabilidades, $f(t)$, que deve satisfazer as seguintes propriedades:

1. $f(t) \geq 0$,
2. $\int_{-\infty}^{\infty} f(t)dt = 1$,
3. $P(a \leq T \leq b) = \int_a^b f(t)\, dt, \forall\ 0 \leq a \leq b$.

2.3.2 Função de sobrevivência

A função de sobrevivência é dada pela probabilidade do elemento ou indivíduo não falhar até um determinado tempo t. No caso de variáveis aleatórias contínuas, é caracterizada por ser não-crescente, absolutamente contínua, e pode ser escrita da seguinte forma:

$$S(t) = P(T \geq t).$$

Portanto, a função de distribuição de probabilidade acumulada pode ser definida por:

$$F(t) = 1 - S(t), \tag{1}$$

em que $F(t) = \int_{-\infty}^{t} f(u)du$.

2.3.3 Função de risco ou taxa de falha

A taxa de falha em um determinado intervalo $[t_1, t_2)$ é definida como a probabilidade de que o evento de interesse seja experimentado entre os tempos t_1 e t_2 dado que não falhou até t_1, dividida pelo comprimento do intervalo. Logo, a taxa de falha é definida por:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}.$$

Assim, pode-se generalizar a fórmula acima redefinindo o intervalo para $[t, t + \Delta t)$, tendo como resultado:

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Assumindo Δt pequeno, a taxa de falha representa o risco instantâneo do elemento

falhar, dado que sobreviveu até t. A função de risco é importante para descrever uma possível distribuição para o tempo de vida dos indivíduos, mostrando como a taxa de falha instantânea se altera quando o tempo muda.

Desta forma, a função de taxa de falha é descrita como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2}$$

Assim como no caso discreto, a função de taxa de falha para uma variável aleatória contínua é mais informativa do que a própria função de sobrevivência. E, por não ser limitada superiormente, a função de risco também é mais flexível do que a função de sobrevivência.

2.3.4 Função de risco acumulado

A função de risco acumulada fornece a taxa de risco acumulada do elemento e é denotada por:

$$H(t) = \int_0^t h(u)du.$$

Essa função não possui uma interpretação direta, porém é útil na análise da função de risco. Isso geralmente ocorre em estimações não-paramétricas, onde $H(t)$ apresenta um estimador com propriedades ótimas e $h(t)$ pode ser difícil de ser estimada.

2.4 Relações importantes entre as funções utilizadas em Análise de Sobrevivência

Segundo Nakano (2017), dentro das funções chaves da análise de sobrevivência existem relações matemáticas, que devem ser levadas em consideração por possuírem papel importante na estimação da função de sobrevivência e das funções subsequentes, pois, em alguns casos, é necessário obter essas funções e nem sempre todas as informações sobre as mesmas são conhecidas.

Utilizando as informações dos subitens anteriores, é sabido que a função de sobrevivência corresponde à probabilidade de um elemento não falhar até um tempo t, e também que a função de risco corresponde à probabilidade condicional de o elemento falhar no instante t dado que sobreviveu até esse tempo. Dito isto, a seguinte relação pode ser explicada:

$$h(t) = P(T = t \mid T \geq t) = \frac{P(T = t \mid T \geq t)}{P(T \geq t)} = \frac{P(T = t)}{P(T = t) + P(T \geq t)} = \frac{p(t)}{p(t) + S(t)}.$$

Dessa forma, a função de distribuição de probabilidades, definida na seção 2.3.1 resulta em:

$$p(t) = \frac{h(t)}{1-h(t)} S(t),\ t = 0, 1, 2, ... \tag{3}$$

Portanto, a função de distribuição de probabilidades poderá ser escrita em termos da função de sobrevivência, como segue abaixo:

$$p(t) = \begin{cases} 1 - S(0) & ,\ t = 0 \\ S(t-1) - S(t) & ,\ t = 1, 2, ... \end{cases}$$

Ainda, se $t = 1, 2, ...,$

$$S(t) = \frac{S(0)}{1}\frac{S(1)}{S(0)}\frac{S(2)}{S(1)}...\frac{S(t-1)}{S(t-2)}\frac{S(t)}{S(t-1)} = S(0)\prod_{k}^{t}\frac{S(k)}{S(k-1)}.$$

Sabendo que $S(0) = 1 - p(0)$ e que:

$$h(0) = P(T = 0 \mid T \geq 0) = \frac{P(T = 0 \bigcap T \geq 0)}{P(T \geq 0)} = \frac{P(T = 0)}{P(T \geq 0} = p(0).$$

Tem-se que a função de sobrevivência pode também ser obtida através da função de risco, como descrito abaixo:

$$S(t) = [1 - h(0)]\prod_{k=0}^{t}\frac{S(k)}{p(k) + S(k)} = [1 - h(0)]\prod_{k=0}^{t}\left[1 - \frac{p(k)}{p(k)+S(k)}\right]$$

$$S(t) = [1 - h(0)]\prod_{k=0}^{t}[1 - h(k)],$$

$$S(t) = \prod_{k=0}^{t}[1 - h(k)],\ t = 0, 1, 2, ... \tag{4}$$

Das expressões (3) e (4), obtém-se a distribuição de probabilidades em termos da função de risco, como segue:

$$p(t) = \begin{cases} h(0) & ,\ se\ t = 0 \\ h(t)\prod_{k}^{t-1}[1 - h(k))] & ,\ se\ t = 1, 2, ... \end{cases}$$

As relações expressas acima estão definidas apenas para variáveis aleatórias discretas. O caso de variáveis aleatórias contínuas é análogo. Usando a definição de taxa de

falha descrita em (2), é possível visualizar que:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \bigcap T \geq t)}{\Delta t P(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T > t)}, \end{aligned}$$

$$h(t) = \frac{f(t)}{S(t)}. \tag{5}$$

Portanto, de (5), nota-se que

$$f(t) = h(t)S(t). \tag{6}$$

Sabendo que a função densidade $f(t)$ é definida pela derivada da função de distribuição acumulada, $F(t)$, e que, através de (1), $F(t) = 1 - S(t)$, então:

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t). \tag{7}$$

Se (7) for substituído em (5), tem-se que

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\ log(S(t)). \tag{8}$$

Integrando os dois lados de (8):

$$\log S(t) = -\int_0^t h(u)du = -H(t),$$

Implica em:

$$S(t) = \exp - \int_0^t h(u)du = \exp[-H(t)]. \tag{9}$$

Por fim, substituindo (9) na equação (6), obtém-se:

$$f(t) = h(t) \exp[-H(t)].$$

Esta última expressão é útil para desenvolver procedimentos de estimação baseados apenas nas funções de risco.

2.5 Estimação não-paramétrica

Por mais complexo que seja o estudo, é crucial que o primeiro passo a ser dado, para melhor entendimento das variáveis de interesse, seja a análise descritiva dos dados. Em sobrevivência, a presença de observações censuradas é um problema para o uso de técnicas convencionais de análise descritiva, como por exemplo, histogramas, médias, box plot entre outros.

Nos textos base de estatística, a análise descritiva consiste em encontrar medidas de tendência central e variabilidade. Como a presença de censura invalida esse tratamento utilizado para os dados, a análise descritiva envolvendo os dados do tempo de vida de determinado elemento é realizada pela função de sobrevivência (Colosimo; Giolo, 2009).

Neste caso, estimar a função de sobrevivência é útil tanto como forma de descrição dos dados observados, quanto para estimação de quantidades importantes que serão utilizadas posteriormente, como, por exemplo, a função de risco.

Na ausência de censura, a estimação da função de sobrevivência pode ser efetuada de forma empírica, uma vez que corresponde à probabilidade de dado elemento não falhar em um tempo t, e pode ser descrita da seguinte forma:

$$\hat{S(t)} = \frac{número\ de\ elementos\ que\ não\ falharam\ até\ o\ tempo\ t = t_0}{número\ de\ elementos\ no\ estudo}. \qquad (10)$$

Na prática, os dados observados em análise de sobrevivência possuem censuras, o que requer um cuidado maior e técnicas especializadas para agregar essas informações ao estudo. Como definido anteriormente, a censura indica que o tempo de falha de um determinado elemento foi maior do que o tempo registrado.

Para estimação da função de sobrevivência, há três mecanismos mais conhecidos: Estimador de Kaplan-Meier, Estimador de Nelson-Aalen e Tábuas de vida. As tábuas de vida foram muito utilizadas por volta do século XIX e serviam, essencialmente, para grandes amostras, para estimar características associadas ao tempo de vida dos seres humanos. O estimador de Nelson- Aalen apresenta, essencialmente as mesmas características do estimador de Kaplan-Meier.

Neste trabalho, será utilizado o estimador de Kaplan-Meier, que é, sem dúvida, o mais utilizado na atualidade e que vem ganhando cada vez mais espaço em estudos de análise de sobrevivência.

Segundo Colosimo e Giolo, o estimador não-paramétrico de Kaplan-Meier, proposto por Kaplan e Meier(1958) para estimar a função de sobrevivência, é também chamado de

estimador limite-produto. Ele é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como na equação (10).

Dessa maneira, $\hat{S(t)}$ assume a forma de uma função escada, que possui degraus para cada tempo observado de falha, cujo tamanho corresponde ao número de falhas em cada tempo t. O estimador de Kaplan-Meier utiliza quantos intervalos forem necessários para cada tempo de falha distinta.

Para definir o estimador de Kaplan-Meier é necessário considerar as seguintes suposições:

- $t_1 < t_2 < t_3 < ... < t_k$ k tempos ordenados e distintos de falha.
- Seja d_i o número de falhas, com $i = 1, 2, ..., k$, e
- n_i o número de indivíduos sob risco em t_i, ou seja, os indivíduos que não experimentaram o evento de interesse ou que não foram censurados até o tempo t_i.

Assim, é possível definir o estimador como:

$$\hat{S}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i} \right).$$

Uma das razões da adequação do estimador de Kaplan-Meier à estimação não-paramétrica da função de sobrevivência é o fato de que ele se apresenta como um estimador de máxima verossimilhança da função de sobrevivência, como foi justificado por Kaplan e Meier em seu artigo original. Segundo Colosimo e Giolo (2006), as propriedades do estimador de Kaplan-Meier são as seguintes:

- Não-viciado para amostras grandes;
- Fracamente consistente;
- Converge assintoticamente para um processo Gaussiano, e
- É estimador de máxima verossimilhança de S(t).

A consistência e a normalidade assintótica de $\hat{S}(t)$ foram provadas, sob certas condições de regularidade, por Breslow e Crowley(1974) e Meier (1975), e, como dito anteriormente, no artigo original, Kaplan e Meier(1958) mostram que $\hat{S}(t)$ é estimador de máxima verossimilhança de $S(t)$ (Colosimo;Giolo, 2006).

Utilizando a propriedade assintótica do estimador, é possível definir sua variância assintótica e assim, um intervalo de confiança para a função de sobrevivência. A variância

é estimada pela fórmula de Greenwood e é expressa por:

$$\widehat{Var}(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \sum_{i:t_i<t} \frac{d_i}{n_i(n_i - d_i)}.$$

Sabendo que a estimativa de $S(t)$ tem distribuição assintótica Normal, o intervalo aproximado de $100(1-\alpha)\%$ de confiança para S(t) é dado por:

$$I.C.[S(t); 100(1-\alpha)\%] : \hat{S}(t) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{S}(t))}.$$

É importante salientar que, para valores extremos de t, o intervalo de confiança pode apresentar valores negativos no limite inferior e valores maiores que 1 no limite superior. Como a função de sobrevivência é limitada no intervalo $[0, 1]$, então é necessário realizar um truncamento no resultado. Para resolver esse problema, uma solução adequada seria utilizar o intervalo de confiança "Log-log", proposto por Kalbfleish e Prentice(2002), em que a estimativa da função de sobrevivência passa a ser $\widehat{U(t)} = \log[-\log(\hat{S(t)})]$. Assim, a variância assintótica estimada passa a ser:

$$\widehat{Var}(\widehat{U}(t)) = \frac{\sum_{i:t_i<t} \frac{d_i}{n_i(n_i-d_i)}}{\left[\sum_{i:t_i<t} \log\left(\frac{d_i}{n_i(n_i-d_i)}\right)\right]^2} = \frac{\sum_{i:t_i<t} \frac{d_i}{n_i(n_i-d_i)}}{\left[\log\ \hat{S}(t)\right]^2}.$$

Logo, o intervalo aproximado de $100(1-\alpha)\%$ de confiança para S(t), ao utilizar a variância assintótica estimada, é definido por:

$$I.C.[S(t); 100(1-\alpha)\%] : [\hat{S}(t)]^{\exp\left(\pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{U}(t))}\right)}.$$

2.6 Seleção de modelos probabilísticos

Como citado anteriormente na seção 2.3.3, a função de risco agrega mais informação do que a função de sobrevivência, uma vez que a estimação da função de risco pode trazer indícios de qual modelo probabilístico se adequaria melhor aos dados do estudo em questão. Há duas formas importantes de se avaliar qual modelo pode ser um bom candidato: A curva do Tempo Total em Teste (TTT) e o gráfico da função de risco acumulado, que serão explicados de forma mais detalhada a seguir:

2.6.1 Curva TTT

Uma forma simples de seleção de possíveis modelos probabilísticos é através do gráfico do Tempo Total em Teste (Curva TTT). Esse método foi proposto por Aarset em 1987 e é definido da seguinte forma:

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{n} T_i}, \tag{11}$$

Em que $r = 1, 2, ..., n$ e $T_{i:n}, i = 1, 2, ..., n$ são estatísticas de ordem da amostra (ordenada de forma crescente).

A Figura 1 mostra várias formas cuja função $G(r/n)$, definida na equação (11) pode assumir. A interpretação da Figura 1 é dada por:

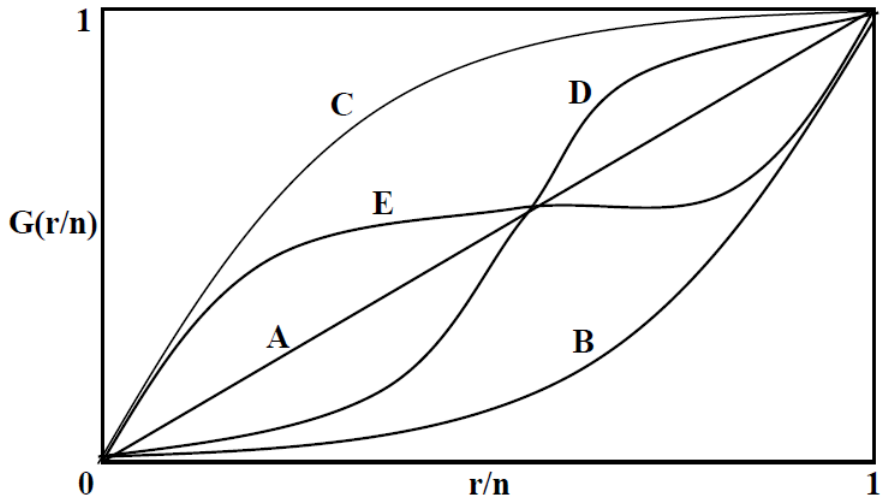

Figura 1: Possíveis curvas no TTT_plot

- Se o gráfico apresenta uma reta diagonal (curva A), então a função de risco é constante. Assim, o modelo Exponencial seria um bom candidato para o ajuste do modelo.

- Caso apresente uma curva côncava (curva C) ou convexa (curva B), então a função de risco é monotonicamente crescente ou decrescente, respectivamente. Um modelo a ser considerado nesse caso seria o Weibull.

- Se a curva TTT apresentar um aspecto côncavo e logo após convexo (curva E), então a função de risco é unimodal e os modelos Log-Normal, Log-Logístico e Burr XII poderiam ser adequados. No caso contrário, ou seja, se a curva for convexa e após côncava (curva D), então a função de risco apresenta um formato de "U",

e as distribuições Weibull modificada generalizada e Kumaraswamy Generalizada poderiam ser candidatas adequadas.

- Se o gráfico apresentar várias regiões côncavas e convexas, então a função possui um risco multimodal e , logo, distribuições de misturas ou modelos de riscos múltiplos podem ser considerados.

Um aspecto relevante da curva TTT é o fato que a sua formulação não considera a existência de censuras. Portanto, é necessário ter cuidado ao optar por essa opção de avaliação da função de risco, uma vez que ela pode acarretar erros de interpretação na seleção dos modelos, no caso de um estudo que possua muitas censuras.

2.6.2 Gráfico de $\hat{H}(t)$

Como a curva TTT não considera a presença de censuras, uma alternativa adequada é a construção do gráfico da estimativa da função de risco acumulado (ou taxa de falha acumulada), $\hat{H}(t)$, cujo comportamento se assemelha ao de $H(t)$. Isto posto, a forma que a função de risco acumulado assumir poderá trazer indícios importantes de qual modelo probabilístico se ajusta melhor aos dados estudados.

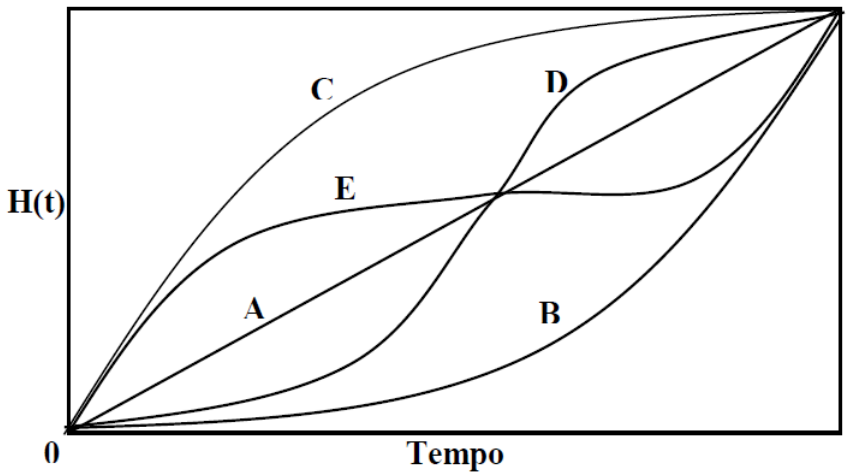

Figura 2: Possíveis curvas no gráfico da função de risco acumulado

Na Figura 2 há exemplos de formas comuns que a função $\hat{H}(t)$ pode assumir, e sua interpretação é dada por:

- Se $\hat{H}(t)$ possuir a forma de uma reta diagonal (curva A), então a função de risco é constante e um bom modelo seria o Exponencial, assim como na curva TTT.

- Caso o gráfico apresente uma curva convexa (curva B) ou côncava (curva C), a função de risco é monotonicamente crescente ou decrescente, respectivamente, e o

modelo Weibull pode ser apropriado.

- Se $\hat{H}(t)$ assumir a forma de uma curva convexa e côncava em seguida (curva D), então o risco apresenta a forma de "U", e no caso contrário (curva E), o risco é unimodal e os modelos Log-Normal, Log-Logístico e Burr XII são bons candidatos.
- Assim como na curva TTT, se o gráfico apresentar várias regiões côncavas e convexas, então o risco é multimodal e os modelos de distribuição mista e riscos múltiplos devem ser considerados.

É importante notar que a interpretação do gráfico da estimativa da função de risco acumulado é o contrário da interpretação da curva TTT, o que requer critério e atenção na hora da avaliação dos mesmos.

2.7 Distribuição Log-Logística

Para uma uma variável aleatória contínua T, a distribuição Log-Logística serve como uma alternativa à distribuição Weibull e sua função densidade de probabilidades é dada por:

$$f(t) = \frac{(\gamma/\alpha)(t/\alpha)^{\gamma} - 1}{[1 + (t/\alpha)^{\gamma}]^2}, t > 0,$$

sendo que $\alpha > 0$ e $\gamma > 0$ são os parâmetros de escala e forma da distribuição, respectivamente. Segundo Lawless (1944), a função de sobrevivência e a função de risco acumulado apresentam formas mais simples e são dadas, respectivamente, por:

$$S(t) = [1 + (t/\alpha)^{\gamma}]^{-}1$$

e

$$h(t) = \frac{(\gamma/\alpha)(t/\alpha)^{\gamma} - 1}{[1 + (t/\alpha)^{\gamma}]}.$$

Uma das vantagens do uso da distribuição Log-Logística é que ela possui uma forma flexível, podendo se adaptar aos dados de forma mais adequada, como demonstra a Figura 3:

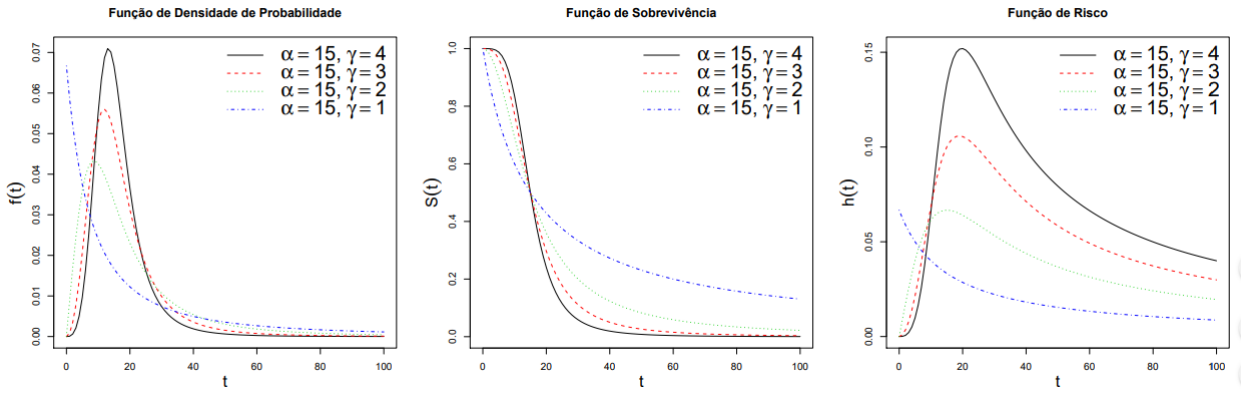

Figura 3: Possíveis formas que a distribuição Log-Logística pode assumir, dependendo dos valores dos parâmetros α e γ

2.8 Discretização de variáveis aleatórias contínuas

Em alguns casos, é válido utilizar distribuições contínuas para gerar modelos discretos análogos. Uma forma simples é agrupar os tempos em intervalos unitários. O método utiliza a "parte inteira"de uma variável aleatória contínua X. Assim, a distribuição de probabilidades da variável aleatória discreta T pode ser escrita da seguinte forma:(Nakano, 2017)

$$p(t) = P(T = t) = P(t \leq X \leq t + 1) = F_X(t + 1) - F_X(t),\ t = 0, 1, 2, ...$$

Portanto, a função de distribuição pode também ser definida em termos da função de sobrevivência, como segue:

$$p(t) = P(T = t) = S_x(t) - S_x(t - 1)$$

2.9 Distribuição Log-Logística discreta

Como será melhor definida na seção de material, a variável resposta (tempo até a adesão da política ZEIS ou Plano Diretor) consiste em uma variável aleatória discreta, e para tanto será necessário realizar a discretização da distribuição Log-Logística, como mostrado na seção 2.8. Dessa forma, a função de distribuição de probabilidades, segundo Santos (2017) é dada por:

$$p(t; \alpha, \gamma) = \frac{1}{1 + (t/\alpha)^\gamma} - \frac{1}{1 + [(t + 1)/\alpha]^\gamma}, t = 0, 1, 2, 3, ...$$

Assim, a função de sobrevivência e de risco são definidas, respectivamente, por:

$$S(t;\alpha,\gamma) = \frac{1}{1+[(t+1)/\alpha]^{\gamma}}, t = 0,1,2,...$$

e

$$h(t;\alpha,\gamma) = 1 - \frac{1+(t/\alpha)^{\gamma}}{1+[(t+1)/\alpha]^{\gamma}}, t = 0,1,2,...$$

sendo que $\alpha > 0$ e $\gamma > 0$ parâmetros de escala e forma, respectivamente.

2.10 Método de Estimação de Máxima Verossimilhança

Para a estimação dos parâmetros da distribuição a partir de dados amostrais, são necessárias técnicas que minimizem o erro. Dentre as mais variadas, o método de máxima verossimilhança é o mais adequado dentro da análise de sobrevivência pois ele incorpora a informação de censura, apresenta boas propriedades assintóticas e possui uma forma "simples".

A ideia básica da estimação por máxima verossimilhança corresponde à escolha de parâmetros que melhor definam a amostra observada. Inicialmente, tem-se que a função de verossimilhança sem considerar as censuras é dada pelo produtório da função densidade de probabilidades, definida por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta),$$

em que θ pode corresponder a um parâmetro ou a um vetor de parâmetros.

Como o objetivo é realizar a estimação dos parâmetros na análise de sobrevivência, é preciso fazer uma adaptação da função $L(\theta)$ de forma que ela apresente também a informação de censura. É sabido que a contribuição de cada informação não censurada é dada pela função densidade, como mostrada na equação anterior. Para os dados censurados, só existe a informação de que o tempo de falha é maior do que o observado, assim, a contribuição desses dados será dada pela função de sobrevivência, pois como mostrado anteriormente nessa revisão, ela é definida como a probabilidade de um elemento não falhar até determinado tempo t. Logo, a função de verossimilhança agora será dada por:

$$L(\theta) = \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(t_i;\theta),$$

sendo que r corresponde ao total de observações não censuradas e $n-r$ corresponde às observações censuradas. Considerando o tipo de censura aleatória e que a censura estudada é um caso particular da mesma, tem-se que a função de máxima verossimilhança

passa a ser proporcional a:

$$L(\theta) \propto \prod_{i=1}^{\infty} [f(t_i;\theta)]^{\delta_i} [S(t_i;\theta)]^{1-\delta_i}, \tag{12}$$

em que δ_i é a variável indicadora de censura. Ao aplicar o logaritmo nos dois lados da equação (12), tem-se, portanto:

$$\log L(\theta) = \sum_{i=1}^{n} \delta_i \log[f(t_i;\theta)] + (1-\delta_i) \log[S(t_i;\theta)]. \tag{13}$$

Como o objetivo desse método é obter parâmetros que maximizem a função de verossimilhança, basta derivar a equação (13) e então os parâmetros estimados serão dados pelos valores que satisfaçam a equação:

$$\frac{\partial}{\partial \theta} \log L(\theta) = 0.$$

3 Metodologia

3.1 Material

Os conjuntos de dados que serão analisados foram disponibilizados por meio de uma parceria com o Instituto de Ciências Políticas da UnB e contém informações sobre os 5570 municípios brasileiros. O período analisado para este estudo será de 1997 a 2015, considerando uma importante variável no estudo, chamada "Competição Política", que passou a ser registrada a partir das eleições de 1996. Portanto, o tempo inicial será $t_i = 1997$, pois dessa forma é possível avaliar o impacto dessa variável no tempo de adesão do município à política de interesse (ZEIS, Plano Diretor).

É importante conceituar essas políticas de interesse, a fim de facilitar o entendimento do estudo. A política Plano Diretor é o instrumento básico para o desenvolvimento de um município, instituído pela Constituição Federal de 1988, cujo objetivo principal é "instruir" tanto o poder público quanto a iniciativa privada em relação ao desenvolvimento e crescimento urbano e rural, com o intuito de garantir melhores condições de vida para a população.

As Zonas Especiais de Interesse Social (ZEIS) surgiram a partir da década de 1980 e são definidas como a parcela de área urbana instituída pelo Plano Diretor ou definida por outra lei municipal, destinada preponderantemente à população de baixa renda através de:

- urbanização de bairros ou imóveis públicos;
- aprovação de loteamentos ou desmembramentos;
- regularização de núcleos urbanos informais consolidados.

A análise é realizada através de duas variáveis respostas: o tempo até a adesão da política ZEIS via legislação específica e o tempo até a adoção da ZEIS independente do tipo, se a adoção é realizada por meio da política Plano Diretor ou via legislação específica. Como as duas variáveis são medidas em anos completos, o estudo utilizará modelos para variáveis aleatórias discretas.

Da verificação preliminar dos bancos de dados, foi possível notar que para alguns municípios, o ano de adoção da política ZEIS apresenta valores como "Não soube informar" e "Recusa". Sendo assim, optou-se pela exclusão desses dados do banco para evitar possíveis erros na análise. Após a exclusão dos municípios que falharam antes do ano de 1997, o banco remanescente ficou com 5445 municípios.

O banco de dados da política ZEIS possui as seguintes variáveis:

- Código do município;
- Adoção da política ZEIS via legislação específica;
- Tempo até a adoção da política via legislação específica;
- Adoção da política ZEIS independente do tipo;
- Tempo até a adoção da política ZEIS independente do tipo;

Para verificar se há a influência de outros fatores no tempo de adesão à política ZEIS, foram incluídas dez covariáveis, que representavam aspectos políticos e geográficos dos municípios. A inclusão foi realizada através da correspondência entre o período no qual foram medidas essas covariáveis e o ano de falha, isto é, o município que falhou recebe as informações referentes ao período imediatamente anterior ou igual ao ano em que aderiu a ZEIS.

- Margem de vitória: apresenta o percentual de vitória do candidato eleito em relação ao segundo colocado, medido a partir de 1996 e com periodicidade de quatro em quatro anos, obrigatoriamente em anos eleitorais;
- NEP: número efetivo de partidos políticos, medido a partir do ano 2000 e com periodicidade de quatro em quatro anos, obrigatoriamente em anos eleitorais;
- Região: é representada por 4 variáveis *dummies*, em que a categoria de referência será a Região Nordeste;
- Conselho de Política Urbana: variável binária, que indica se foi criado no município um conselho de políticas urbanas no período anterior à falha do município, com medição a partir do momento da sua criação independente da existência de período eleitoral;
- População: devido ao tamanho dos municípios, à diferença entre os valores correspondentes ao tamanho da sua população e as demais variáveis, será considerado o logaritmo da população de cada município, obtida por meio do censo demográfico de 2000 e 2010;
- Prefeito Reeleito: variável binária que indica se o prefeito foi reeleito no ano anterior ou igual ao que o município falhou, medida em anos eleitorais a partir de 1996;
- ano eleitoral: variável binária que indica se o ano de falha também ano eleitoral no município.

3.2 Métodos

3.2.1 Modelo de Regressão Log-Logística

Usualmente, quando se verifica que algumas variáveis de interesse podem estar correlacionadas com o tempo de falha, essa relação pode ser expressa através de modelos de regressão, que podem ser obtidos por diversas formas na análise de sobrevivência, entre elas:

- Modelos obtidos por meio de reparametrização da distribuição de probabilidades
- Modelos de locação e escala
- Modelo de riscos proporcionais - semi-paramétrico

Dentre os modelos mencionados, o estudo utilizará os modelos obtidos por meio da reparametrização da distribuição de probabilidades.

Ao realizar a reparametrização da distribuição da variável resposta, esta pode se conectar às variáveis explicativas através de uma função de ligação $g(.)$, que segundo Santos (2017), é dada por:

$$\theta = g(\eta) \tag{14}$$

Com $\eta = \mathbf{x}^T\beta$ correspondendo ao preditor da variável tempo de falha, $\beta = (\beta_0, ..., \beta_p)^T$ é o vetor dos coeficientes da regressão e $\mathbf{x^T} = (1, x_1, ..., x_p)$ o vetor de covariáveis.

Há diversos tipos de funções de ligação e como a reparametrização será realizada no parâmetro alpha que é positivo, será utilizada a função $\alpha = g(\eta) = \exp(\eta) = \exp(\mathbf{x}^T\beta)$.

Ao considerar a distribuição log-logística discreta definida na seção 2.9, tem-se que:

$$p(t; \alpha, \gamma) = \frac{1}{1 + (t/\alpha)^\gamma} - \frac{1}{1 + [(t+1)/\alpha]^\gamma}, t = 0, 1, 2, 3, ... \tag{15}$$

Como α é o parâmetro de escala da função e é positivo, o modelo de regressão Log-Logístico será definido por:

$$p(t; \beta, \gamma) = \frac{1}{1 + (t/\exp(\mathbf{x^T}\beta))^\gamma} - \frac{1}{1 + [(t+1)/\exp(\mathbf{x^T}\beta)]^\gamma}, t = 0, 1, 2, 3, ... \tag{16}$$

Consequentemente, a função de sobrevivência e a função de risco serão dadas por:

$$S(t; \beta, \gamma) = \frac{1}{1 + [(t+1)/\exp(\mathbf{x^T}\beta]^\gamma}, t = 0, 1, 2, ... \tag{17}$$

$$h(t;\beta,\gamma) = 1 - \frac{1+(t/\exp(\mathbf{x}\beta))^{\gamma}}{1+[(t+1)/\exp(\mathbf{x}\beta)]^{\gamma}}, t = 0,1,2,... \tag{18}$$

Dada a função de sobrevivência, a interpretação dos coeficientes de regressão é definida por meio da função quantil da distribuição Log-Logística discreta. Uma proposta para interpretação pode ser obtida através da razão de tempos medianos (Santos,2017), isto é $q_{0.5}$. Essa razão é definida como:

$$t_{0,5}(\hat{\alpha};\hat{\gamma}) = inf\left\{t : \hat{\alpha}\left[\frac{0,5}{(1-0,5)}\right]^{\frac{1}{\hat{\gamma}}} - 1 \leq t\right\} \cong \hat{\alpha} - 1, \tag{19}$$

considerando $\hat{\alpha} = \exp(\mathbf{x^T}\hat{\beta})$, tem-se então que $t_{0,5} + 1 = \exp(\mathbf{x^T}\hat{\beta})$. Logo, a razão dos tempos medianos do modelo de regressão Log-Logístico discreto para uma covariável dicotômica é dada pela expressão:

$$\frac{1+t_{0,5}(x=1,\hat{\gamma},\hat{\beta})}{1+t_{0,5}(x=0,\hat{\gamma},\hat{\beta})} = \frac{\exp(\hat{\beta}_{\mathbf{0}} + \hat{\beta}_{\mathbf{1}}\mathbf{x})}{\exp(\hat{\beta}_0)} = \exp^{\beta_{\mathbf{1}}\mathbf{x}} \tag{20}$$

Portanto, segundo Santos (2017), se o coeficiente de regressão for positivo, interpreta-se que o tempo mediano mais uma unidade ($t_{0,5+1}$) de um município que possui determinada característica dada como $x = 1$ é $\exp^{\hat{\beta}_1}$ vezes o tempo mediano mais um ano de um município que não possui a característica. Nos casos em que o coeficiente de regressão é negativo conclui-se que o tempo mediano de sobrevivência mais um ano é $\exp^{\hat{\beta}_p}$ vezes menor para o município que tem a característica dada como $x = 1$ em relação àqueles que não a possuem.

Assim, utilizando o método de estimação de máxima verossimilhança descrito na seção 2.10, os parâmetros do modelo de regressão Log-Logístico discreto serão estimados através da maximização da seguinte equação:

$$\log L(\beta,\gamma) = \sum_{i=1}^{n} \delta_i \log\left[\frac{1}{1+(\frac{t}{\exp(\mathbf{x^T}\beta)})^{\gamma}} - \frac{1}{1+[\frac{(t+1)}{\exp(\mathbf{x^T}\beta)}/]^{\gamma}}\right] + (1-\delta_i)\log\left[\frac{1}{1+[\frac{(t+1)}{\exp(\mathbf{x^T}\beta)}]^{\gamma}}\right] \tag{21}$$

3.2.2 Resíduos de Cox-Snell

Após a estimação dos parâmetros do modelo, a análise de resíduos de Cox-Snell é uma ferramenta útil para verificação da qualidade do ajuste. Os resíduos são quantidades

determinadas da seguinte forma (COLOSIMO;Giolo, 2006):

$$\hat{e_i} = \Lambda(t_i \mid \mathbf{x_i}),$$

em que $\Lambda(.)$ é a função de taxa de falha acumulada obtida do modelo ajustado. Os resíduos $\hat{e_i}$ vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão se o modelo for adequado aos dados. A função taxa de falha é determinada por $\hat{e_i} = -\log(\hat{S}(\hat{e_i})$. O gráfico de Kaplan-Meier da curva de sobrevivência desses resíduos e da curva obtida pela distribuição exponencial padrão auxiliam na verificação do ajuste do modelo. Quanto mais próxima elas se apresentarem, melhor é considerado o ajuste do modelo aos dados.

4 Resultados e Discussões

4.1 Adesão à política ZEIS via legislação específica

4.1.1 Análise Descritiva

Nesta seção são apresentados os resultados da análise descritiva preliminar da adesão dos municípios à política ZEIS através de lei municipal específica.

A Figura 4 corresponde ao gráfico de Kaplan-Meier para estimação da função de sobrevivência. Nota-se que houve pouca adesão dos municípios brasileiros, (apenas cerca de 10% dos municípios falharam), nota-se também que a estimativa da função de sobrevivência decresce conforme os anos passam, de acordo com o esperado. Assim, a estimativa da função de sobrevivência para o último ano da pesquisa chega a $0,883$.

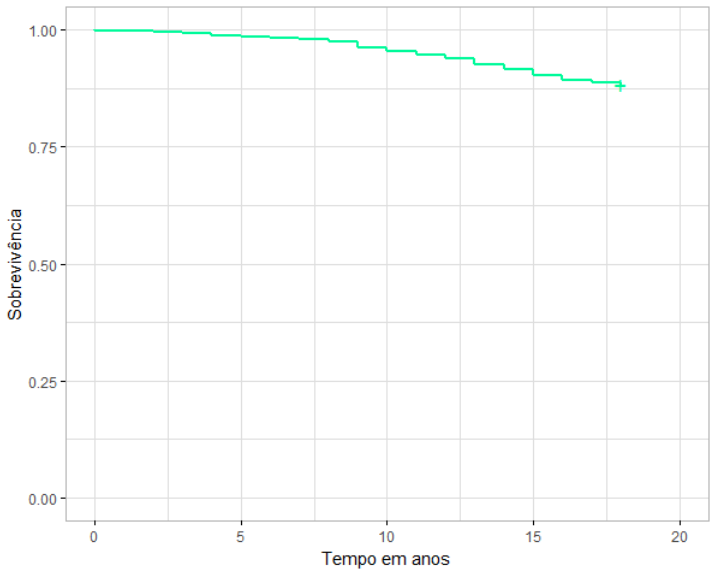

Figura 4: Estimativa da função de sobrevivência para o tempo de adesão dos municípios à política ZEIS via legislação específica

A análise dos gráficos do TTT plot e da função de risco acumulado do tempo até a adesão da política ZEIS via lei municipal específica(Figuras 5 e 6), apresentam indícios de uma função de risco monotonicamente crescente devido ao seu formato notadamente côncavo para o TTT plot e convexo para o gráfico da função de risco acumulado. Logo, a distribuição Weibull discreta pode ser adequada. Como alternativa, pode ser usada a distribuição Log-Logística discreta, pois apesar de possuir risco unimodal, ela permite uma flexibilização maior e pode representar bem os dados.

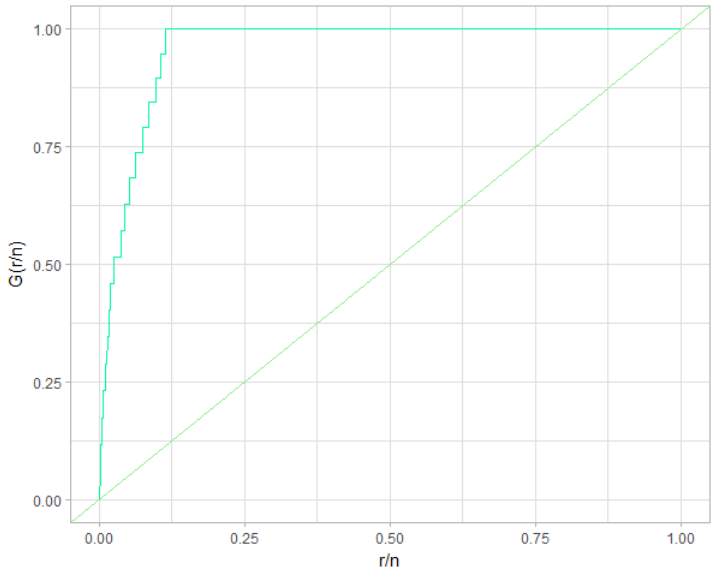

Figura 5: TTT Plot do tempo de adesão à política ZEIS via legislação específica

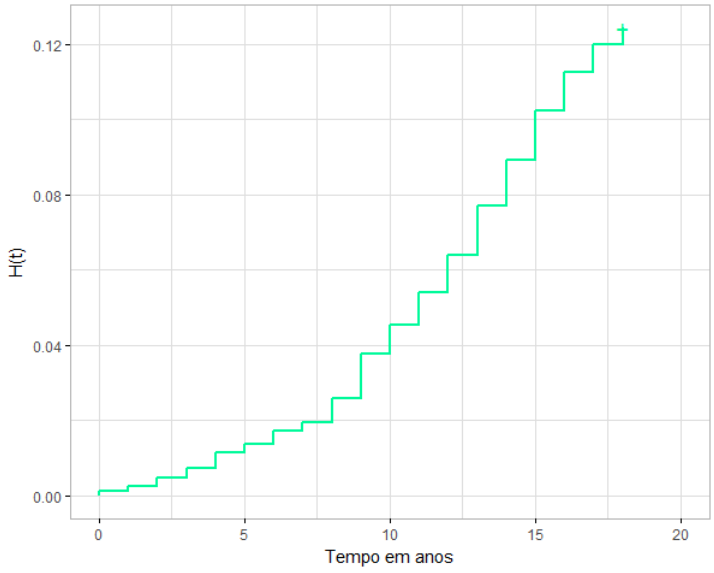

Figura 6: Gráfico da função de risco acumulado do tempo de adesão à política ZEIS via legislação específica

Neste momento, serão realizadas as análises descritivas das covariáveis que poderão ser incluídas no modelo, explicitadas na seção de material.

As primeiras covariáveis consideradas serão margem de vitória, NEP e logaritmo da população. Devido ao tempo de sobrevivência desses dados ser discreto, não é viável realizar a análise dos gráficos de dispersão das mesmas. Em relação à variável região geográfica, nota-se pouca diferença entre as curvas de sobrevivência.

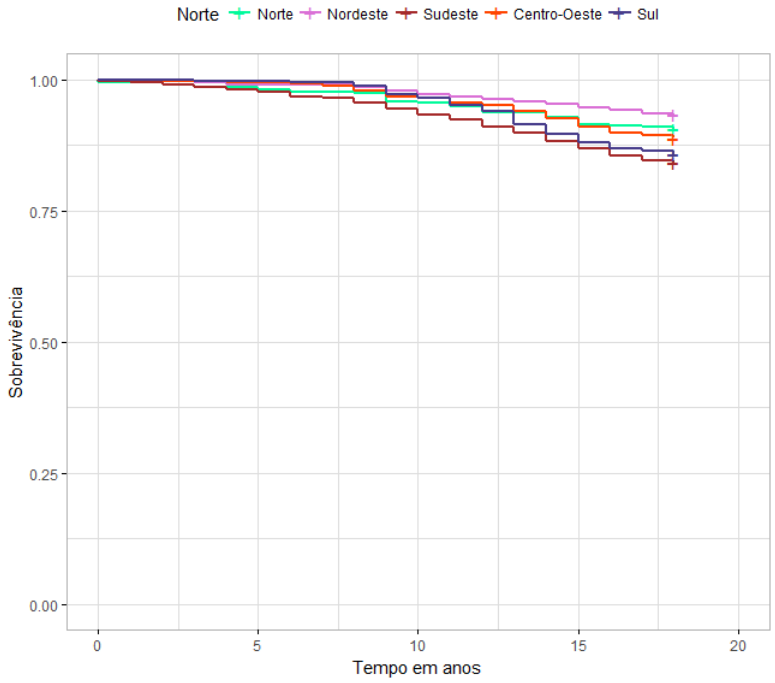

Figura 7: Estimativa de Kaplan-Meier considerando as regiões geográficas brasileiras

Para as variáveis de ano eleitoral e conselho de política urbana, verifica-se que há indícios de diferença entre as curvas de sobrevivência. Entretanto, em relação à variável prefeito reeleito, a diferença entre as curvas é mínima. Portanto, há indícios de que a variável prefeito reeleito não apresente significância estatística no modelo probabilístico de regressão.

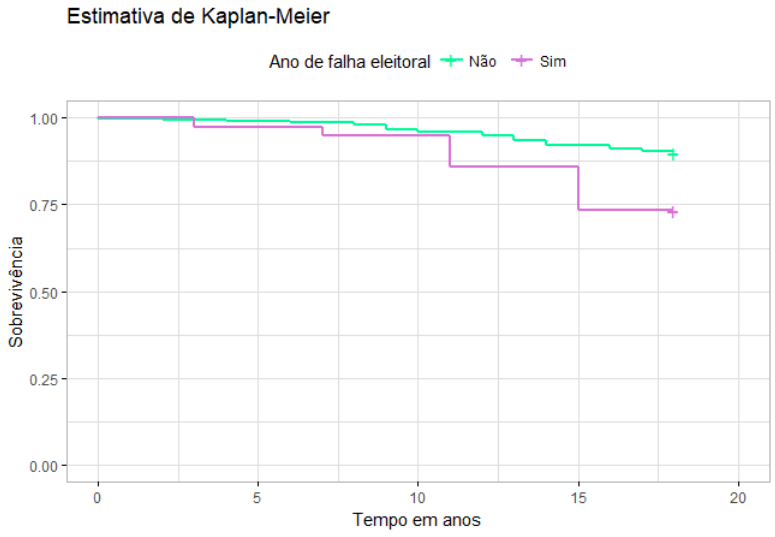

Figura 8: Estimativa de Kaplan-Meier das curvas de sobrevivência considerando se é ano eleitoral

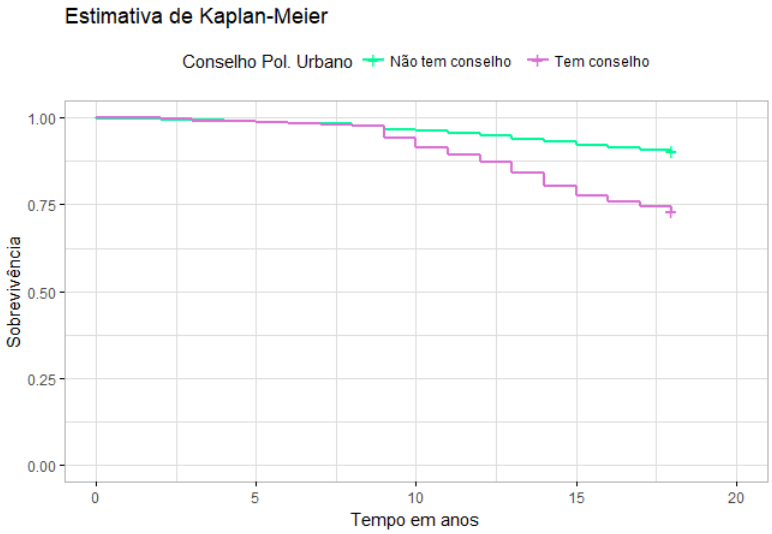

Figura 9: Estimativa de Kaplan-Meier das curvas de sobrevivência considerando se o há conselho de política urbana

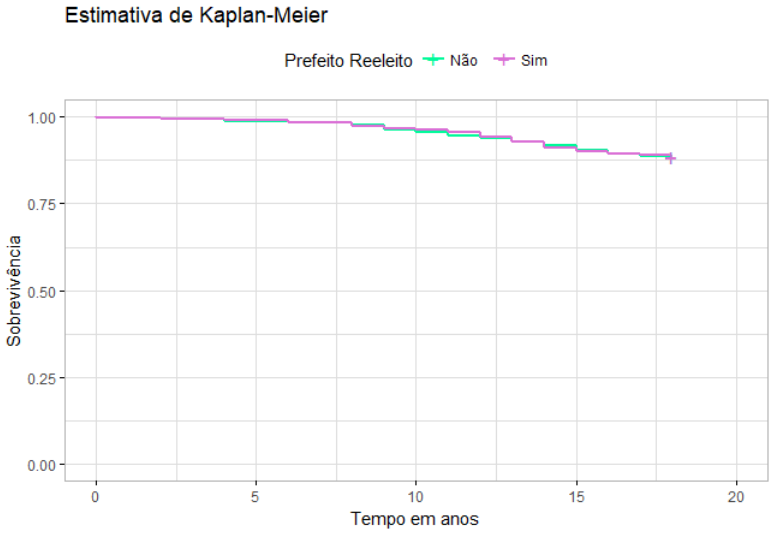

Figura 10: Estimativa de Kaplan-Meier das curvas de sobrevivência considerando se o prefeito foi reeleito

4.1.2 Modelagem

Ao realizar o ajuste, não foi possível utilizar a distribuição Weibull por apresentar problemas de convergência, portanto a distribuição Log-Logística foi utilizada. Dessa forma, nota-se que para a adesão da política ZEIS através de lei específica (Figura 14) a distribuição Log-Logística se ajusta melhor do que para a adesão da política ZEIS independente do tipo, conforme será demonstrado na modelagem dos dados de adesão da política sem levar em consideração a forma de adoção.

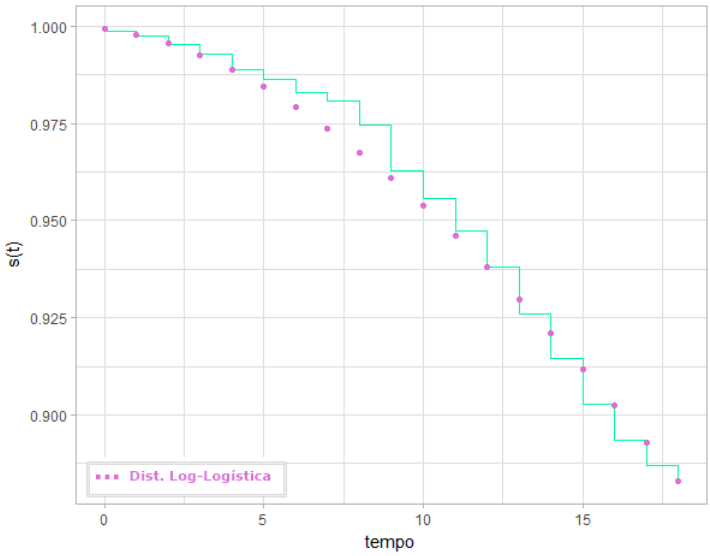

Figura 11: Ajuste da distribuição Log-Logística à função de sobrevivência do tempo de adesão à política ZEIS via legislação específica

Como observado, os dados aparentam se ajustar bem à distribuição Log-Logística discreta e após a estimação dos parâmetros, verifica-se que para α *e* γ os erros-padrão são pequenos. É importante ressaltar que o valor da estimativa de α é bastante elevado, o que demonstra a heterogeneidade dos dados analisados. Portanto se faz necessário um modelo de regressão, para que ao levar em consideração outros fatores que influenciem na adoção da política, seja possível explicar essa heterogeneidade.

Tabela 1: Estimativa dos parâmetros α e β para a distribuição Log-Logística

Parâmetro	Estimativa	Erro Padrão
α	56,067868	2,755329
γ	1,875977	0,07312184

Primeiramente, foram estimados modelos de regressão contendo apenas uma variável explicativa. Para todo esses modelos, os valores das estimativas de β_0 e γ se mantêm em torno de 5,10 e 1,9, respectivamente. A Tabela 2 apresenta os resultados dos coeficientes estimados individualmente.

Tabela 2: Estimativas dos coeficientes para o modelo de regressão Log-Logístico com apenas uma variável

Variável	Estimativa	Erro padrão	P-valor
Margem de Vitória	-0,1177	0,0959	0,2199
NEP	-0,0399	0,0415	0,3372
Cons. Política Urbana	-0,6267	0,0565	<0,0001
log(POP)	-0,1739	0,0191	<0,0001
Prefeito reeleito	-0,0028	0,0491	0,9538
Ano eleitoral	-0,6081	0,0611	<0,0001
NORTE	0,1295	0,0905	0,1522
SUDESTE	-0,2985	0,0474	<0,0001
CENTRO - OESTE	0,0317	0,0830	0,7023
SUL	-0,1496	0,0517	0,0038

Ao considerar um nível de significância de 10% e o p-valor encontrado para cada um dos modelos, percebe-se que apenas as variáveis conselho de política urbana, logaritmo da população, ano eleitoral além de duas *dummies* para as regiões são significativas. Ao utilizar o método *forward* para seleção de variáveis, um modelo final foi alcançado, descrito conforme segue:

$$\alpha = (g(\eta) = \exp \beta_0 + \beta_1 * Cons.Pol.Urbana + \beta_2 * \log(POP) + \beta_3 * Ano.eleitoral) \tag{22}$$

Na Tabela 3 são apresentados os resultados da estimação dos parâmetros do modelo final. É possível perceber que o modelo se ajusta adequadamente aos dados. Além disso, apresentaram um p-valor abaixo do nível de significância de 10%, trazendo evidências que essas variáveis influenciem o tempo de adoção da política ZEIS.

Tabela 3: Estimativas dos parâmetros: modelo final

Parâmetro	Estimativa	Erro Padrão	P - valor
β_0	5,0968	0,2014	<0,0001
β_1	-0,4186	0,0598	<0,0001
β_2	-0,1008	0,0197	<0,0001
β_3	-0,4628	0,0612	<0,0001
γ	1,8966	0,0747	-

Para verificar a qualidade do ajuste, foi realizada uma análise de resíduos de Cox-

Snell. De acordo com o gráfico dos resíduos(Figura 12) há indícios de que o modelo se ajusta razoavelmente bem aos dados, portanto, pode ser considerado para a análise.

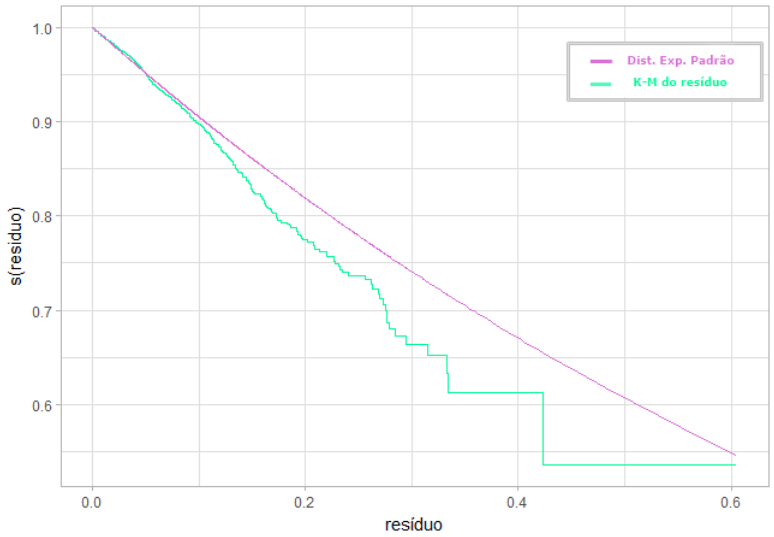

Figura 12: Ajuste dos resíduos Cox-Snell para o modelo selecionado, respectivamente

Através das estimativas definidas na Tabela 3, corroborando o que foi obtido na análise descritiva, o tempo mediano mais um ano dos municípios que não possuíam um conselho de políticas urbanas é $1/\exp(-0,4186) = 1,52$ vezes maior quando comparado aos municípios que possuíam um conselho de políticas urbanas.

Em relação à população, a medida que o logaritmo da população aumenta, a probabilidade de sobrevivência diminui em 0,904 vezes, tudo o mais constante. Também tem-se que o tempo mediano de sobrevivência mais um ano é $1/\exp(-0,4628) = 1,59$ vezes maior para municípios que não estão em ano eleitoral. Em outras palavras, a probabilidade de um município aderir à política ZEIS aumenta se a população é maior, se há um conselho de política urbana e se é ano eleitoral.

4.2 Adoção da ZEIS independente do tipo

4.2.1 Análise descritiva

Nesta seção são apresentados os resultados da análise descritiva preliminar da adesão dos municípios à política ZEIS, sem considerar o tipo de adoção utilizada, via legislação específica ou Plano diretor.

A Figura 13 retrata o gráfico da estimativa da função de sobrevivência por meio do estimador de Kaplan-Meier para a adesão da política ZEIS independente do tipo de adesão. Além disso, são apresentados o gráfico da função de risco e TTT Plot. De acordo

com o gráfico de Kaplan-Meier, é possível perceber que o número de falhas aumenta, chegando a aproximadamente 40% do total. Consequentemente, as estimativas da função de sobrevivência decrescem um pouco mais conforme o passar dos anos.

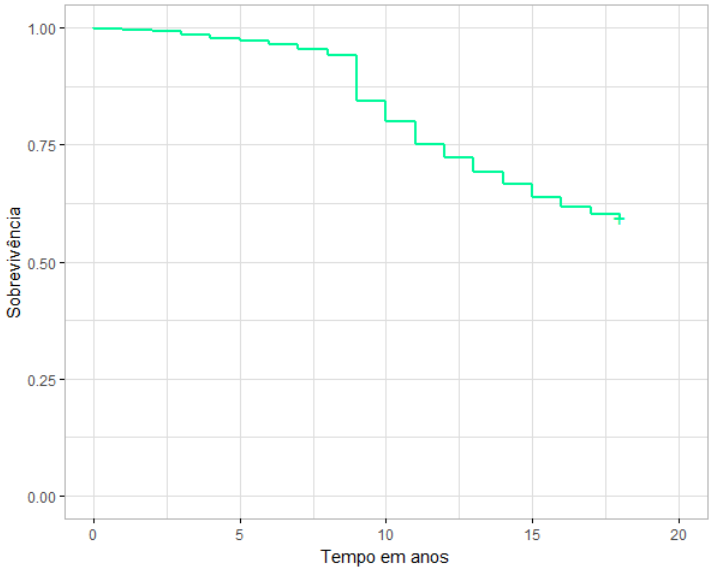

Figura 13: Estimativa da função de sorevivência para o tempo de adesão dos municípios à política ZEIS independente do tipo

Assim como na adesão da política via lei municipal, a análise dos gráficos do TTT Plot e da função de risco acumulado do tempo até a adesão da política ZEIS sem considerar o tipo de adoção(Figuras 14 e 15), apresentam indícios de uma função de risco monotonicamente crescente devido ao seu formato notadamente côncavo no TTT Plot e convexo para o gráfico da função de risco acumulado. Logo, a distribuição Weibull discreta pode ser adequada, assim como a distribuição Log-Logística discreta, pois apesar de possuir risco unimodal, ela permite uma flexibilização maior e pode representar bem os dados.

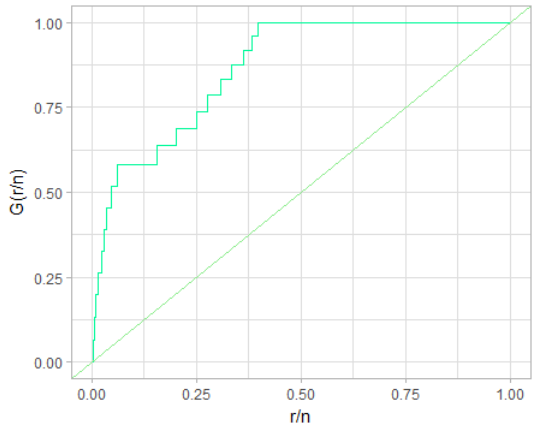

Figura 14: TTT Plot do tempo de adesão à política ZEIS independente do tipo

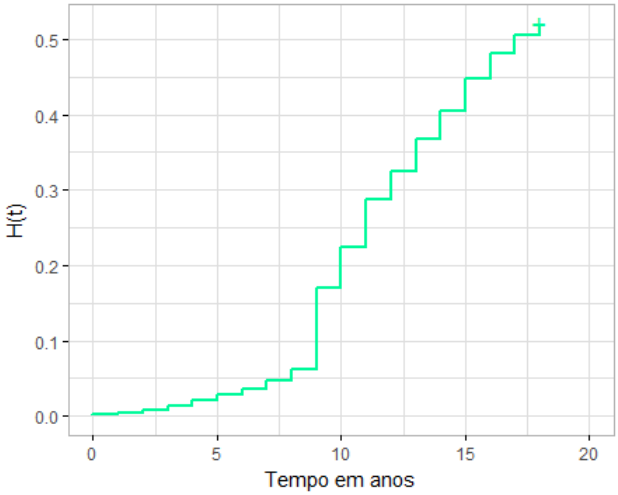

Figura 15: Gráfico da função de risco acumulado do tempo de adesão à política ZEIS independente do tipo

Como as covariáveis serão incluídas no modelo de regressão, são apresentados também os resultados referentes às suas respectivas análises descritivas .

Analisando a variável Região, é notável a diferença entre as Regiões Nordeste e Sul.

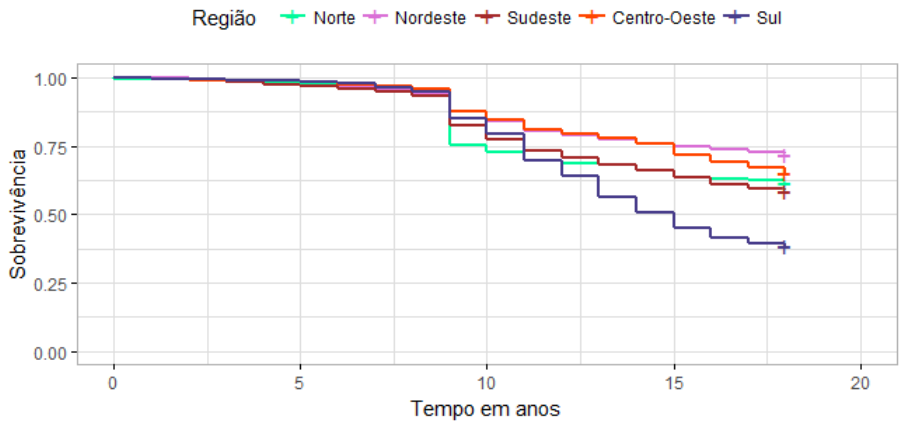

Figura 16: Estimativa de Kaplan-Meier considerando as regiões geográficas brasileiras

Ao avaliar os gráficos de Kaplan- Meier com as variáveis ano eleitoral e conselho de política urbana constata-se que a diferença entre as curvas de sobrevivência possui maior destaque e pode influenciar no modelo de regressão.

Em relação à variável prefeito reeleito, diferente do banco de adoção da ZEIS, há uma diferença mais notável, porém não tão grande entre as curvas de sobrevivência.

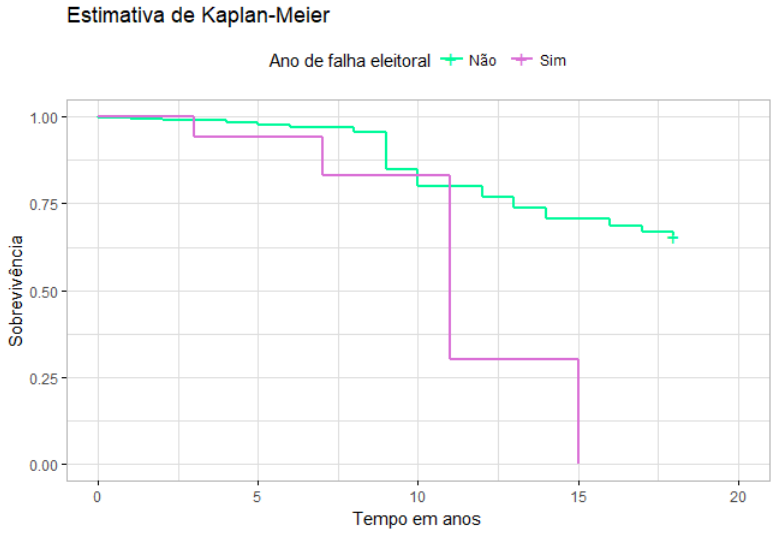

Figura 17: Estimativa de Kaplan-Meier das curvas de sobrevivência considerando se é ano eleitoral

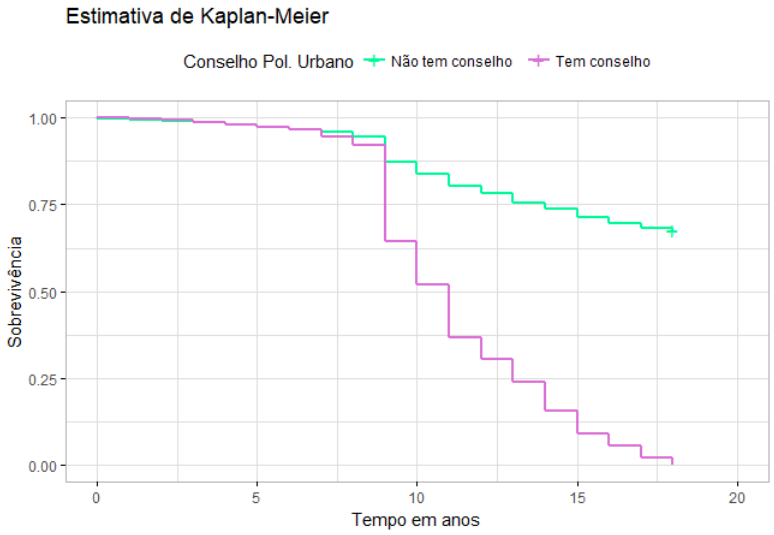

Figura 18: Estimativa de Kaplan-Meier das curvas de sobrevivência considerando se o há conselho de política urbana

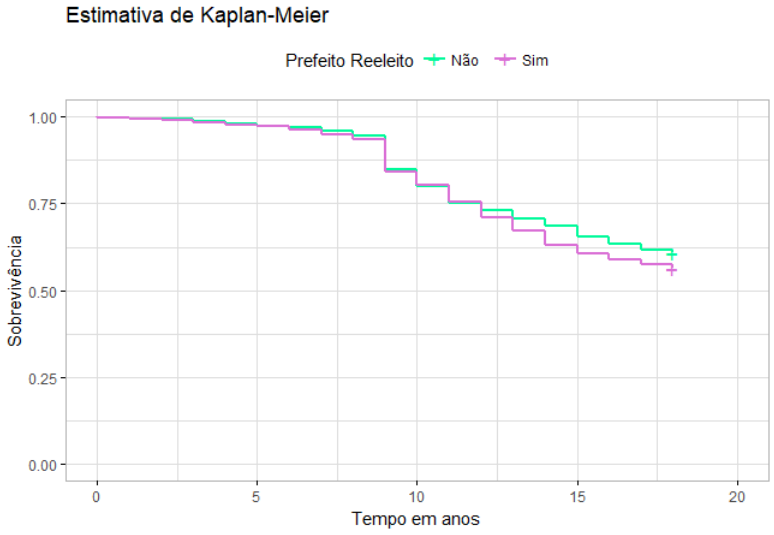

Figura 19: Estimativa de Kaplan-Meier das curvas de sobrevivência considerando se o prefeito foi reeleito

4.2.2 Modelagem

De acordo com a análise descritiva realizada na seção anterior, os gráficos de tempo total em teste (TTT Plot) e do risco acumulado indicaram que a função Weibull discreta poderia se adequar bem aos dados e por possuir risco unimodal e ser flexível, a distibuição Log-Logística discreta pode ser uma alternativa de modelagem. Ao realizar o ajuste, a distribuição Weibull apresentou problemas de convergência em relação à distribuição Log-Logística, por isso, não foi utilizada na modelagem dos dados e na criação do modelo de

regressão.

É possível perceber que a probabilidade da função de sobrevivência estimada pela Log-Logística é subestimada entre os tempos 5 e 9 (aproximadamente) em relação à estimativa não-paramétrica, mas no geral, o ajuste apresenta indícios de adequabilidade.

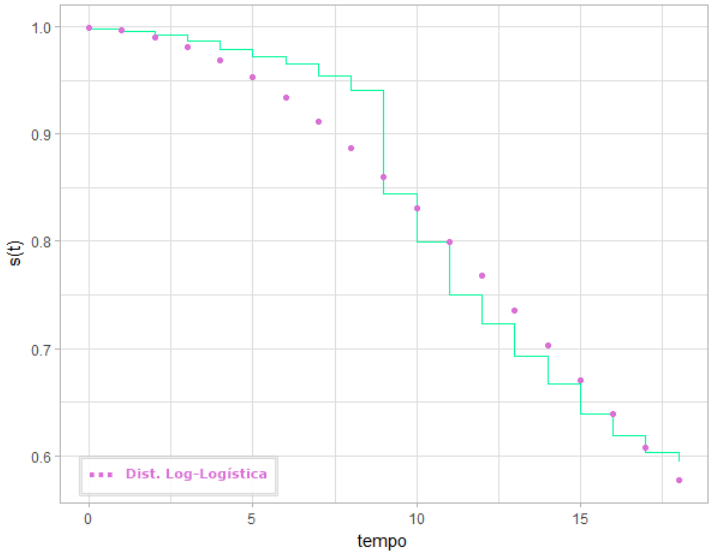

Figura 20: Ajuste da distribuição Log-Logística à função de sobrevivência do tempo de adesão à política ZEIS independente do tipo

Inicialmente, foram obtidas as estimativas dos parâmetros α e γ para um modelo sem covariáveis, através da otimização da função de distribuição explicitada na metodologia. A tabela abaixo apresenta os resultados obtidos com seus respectivos erros padrão:

Tabela 4: Estimativa dos parâmetros α e β para a distribuição Log-Logística

Parâmetro	Estimativa	Erro Padrão
α	21,7309	0,2805
γ	2,3555	0,0453

É perceptível que os erros padrão apresentam valores baixos e assim o modelo parece se ajustar adequadamente aos dados. Após a análise descritiva realizada na seção anterior, todas as variáveis foram inseridas individualmente no modelo de regressão Log-Logístico discreto, através da função de ligação definida na metodologia e foram testadas sob a hipótese nula de não possuírem influência no tempo de adesão, a um nível de 10%. Seguem os resultados obtidos através da otimização da função de regressão:

Tabela 5: Estimativas dos coeficientes para o modelo de regressão Log-Logístico discreto com apenas uma variável

Variável	Estimativa	Erro padrão	P-valor
Margem de Vitória	-0,0988	0,0495	0,0459
NEP	-0,2550	0,0212	<0,0001
Cons. Política Urbana	-0,7487	0,0259	<0,0001
log(POP)	-0,2806	0,0094	<0,0001
Prefeito reeleito	-0,0713	0,0244	0,0035
Ano eleitoral	-0,7154	0,0297	<0,0001
NORTE	-0,0014	0,0425	0,9744
SUDESTE	-0,0356	0,0247	0,1504
CENTRO - OESTE	0,1314	0,0425	0,0020
SUL	-0,3402	0,0251	<0,0001

Para todos os modelos testados, os valores de β_0 e γ tinham suas estimativas em torno de 3,0967 2,5498, com os respectivos erros padrão 0 e 0.0024.

O procedimento *forward* foi utilizado para inserção e seleção das covariáveis que possuem significância no modelo. Antes da escolha do método foi testado um modelo saturado, que apresentou problemas de convergência, e, ao investigar a causa, tem-se que as variáveis do tipo de adoção e de região podem ser apontadas como a fonte, uma vez que não há convergência de nenhum modelo que possua as variáveis supracitadas, apesar de sozinhas funcionarem corretamente.

Inicialmente, para verificar quais variáveis poderiam ser ou não influentes no tempo de adesão à política, foi realizada a modelagem para cada covariável separadamente. Além do problema mencionado, considerando que as Regiões Norte e Sudeste apresentaram indícios de não rejeição da hipótese nula, a retirada dessas variáveis do modelo é uma alternativa viável, buscando minimizar o erro no ajuste e a quantidade de modelos candidatos.

Ao término da seleção, dois modelos finais foram encontrados, sendo o primeiro formado por 4 variáveis explicativas, com o parâmetro α definido como segue:

$$\alpha = g(\eta) = \exp(\beta_0 + \beta_1 * MargemVitoria + \beta_2 * NEP + \beta_3 * Cons.Pol.Urbana + \beta_4 * PreReel) \tag{23}$$

O segundo modelo também dispõe de 4 variáveis, com a diferença de possuir ano eleitoral ao invés de prefeito reeleito, como descrito abaixo:

$$\alpha = \exp(\beta_0 + \beta_1 * Margem_Vitoria + \beta_2 * NEP + \beta_3 * Cons.Pol.Urbana + \beta_4 * Ano.Eleitoral) \quad (24)$$

Com o objetivo de decidir qual é o modelo mais adequado às informações, foram realizados testes de hipóteses para significância dos parâmetros e análise de resíduos. As tabelas 6 e 7 apresentam os resultados obtidos dos dois modelos, juntamente com seus erros padrão.

Tabela 6: Estimativa dos coeficientes do modelo de regressão Log-Logístico discreto para o modelo 2

Parâmetro	Estimativa	Erro	padrão
β_0	3,5542	0,04525	<0,0001
β_1	-0,1499	0,0451	0,0009
β_2	-0,1760	0,0201	<0,0001
β_3	-0,70402	0,0259	<0,0001
β_4	-0,0565	0,0220	0,0103
γ	2,7008	0,0526	-

Tabela 7: Estimativa dos coeficientes de regressão para o modelo 2

Parâmetro	Estimativa	Erro	padrão
β_0	3,5546	0,0452	<0,0001
β_1	-0,1487	0,0451	0,0010
β_2	-0,1764	0,0201	<0,0001
β_3	-0,7039	0,0259	<0,0001
β_4	-0,0564	0,0220	0,0104
γ	2,7005	0,0526	-

É notável a similaridade entre os modelos, tanto para as estimativas, apesar de serem variáveis diferentes, quanto para os erros padrão e para o p-valor. Ao avaliar o ajuste dos resíduos Cox-Snell na Figura 20, não há diferença visível entre os modelos, e o ajuste para a distribuição exponencial está razoável, demonstrando que o modelo log-logístico discreto pode ser aceito como adequado aos dados.

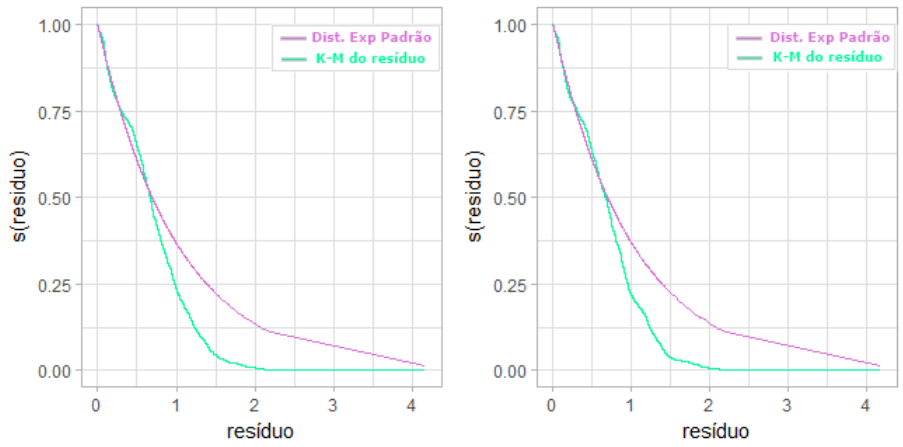

Figura 21: Ajuste dos resíduos Cox-Snell para o modelo 1 e modelo 2, respectivamente

Em relação à interpretação, tem-se que o tempo mediano de sobrevivência mais um ano é 0,861 vezes (considerando os dois modelos) menor quanto maior for a margem de vitória do prefeito eleito em relação ao segundo colocado, tudo o mais constante. O tempo mediano de sobrevivência mais um ano também diminui cerca de 0,838 vezes para cada aumento do número efetivo de partidos vencedores.

Considerando a criação do conselho de política urbana, o tempo mediano de sobrevivência mais um ano é cerca de $1/\exp(-0,4946295) = 1,64$ vezes maior para municípios que não possuem o conselho em relação aos municípios que o possuem. Como exposto na análise descritiva, o tempo mediano de sobrevivência mais um ano, tanto para municípios que não estão em ano eleitoral quanto para municípios que não tiveram o prefeito reeleito, é maior cerca de $1/\exp(-0,945) = 2,57$ vezes do que para o oposto.

Portanto, conclui-se que a probabilidade de um município aderir à política ZEIS independentemente da forma (via legislação específica ou via plano diretor) é maior se houve a criação do conselho de política urbana, se a margem de vitória foi alta, se o número efetivo de partidos for alto, se é ano eleitoral ou se o prefeito foi reeleito. Esses resultados são próximos dos obtidos na análise de adesão da ZEIS via legislação específica.

5 Considerações Finais

O presente trabalho teve como objetivo verificar o efeito das variáveis explicativas geográficas e políticas no tempo de adesão dos municípios à ZEIS. Além disso, foi interessante verificar a melhor forma de medir a adesão dessa política, se via legislação específica ou desconsiderando o tipo de adoção, via Plano Diretor ou via legislação específica.

Foram apresentados os resultados referentes à utilização da distribuição log-logística discreta em dados de análise de sobrevivência e, por meio da inserção de covariáveis, da definição de modelos de regressão discretos que possibilitem melhor compreensão do processo de adesão dos municípios brasileiros à política ZEIS.

Os dados analisados foram provenientes do Instituto de Ciências Políticas - IPOL/UnB e referenciam um estudo real, que aborda diversas políticas públicas e a forma como elas se propagam no país.

Foi realizada uma análise descritiva preliminar. Após, a estimação de um modelo de regressão discreto por meio de algoritmos de otimização da função de máxima verossimilhança. O tempo até a adesão da política ZEIS foi conectado às covariáveis através de uma função de ligação exponencial.

Devido ao método utilizado para obter a função quantil de sobrevivência, a interpretação foi realizada em termos do tempo mediano de sobrevivência, ou seja, da probabilidade de o município aderir à política mais rápido. Conforme a análise descritiva, a interpretação estava de acordo com o esperado.

Ao considerarmos a adoção da política independente do tipo é possível perceber um ganho de informação, pois há um maior número de falhas e consequentemente é possível verificar que mais efeitos influenciam na adoção da ZEIS se comparado aos dados que possuem apenas a adoção via legislação específica.

Posteriormente à realização da análise de resíduos, verificou-se que o modelo teve um ajuste aceitável, portanto, os resultados obtidos poderão contribuir com o Instituto de Ciências Políticas atribuindo uma base estatística aos seus estudos, que apresentam uma fundamentação predominantemente teórica, com extenso potencial na área de análise de sobrevivência.

Como consideração para trabalhos futuros, pode-se cogitar a discretização de outra distribuição probabilística, que seja possivelmente mais flexível que a log-logística discreta, de forma a aumentar a precisão do ajuste.

Referências

Colosimo, E. A. and Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Edgard Blucher, São Paulo. ABE - Projeto Fisher.

Fernandes, L. M. (2013). Inferência bayesiana em modelos discretos com fração de cura. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53:p. 457–481.

Lawless, J. F. (1944). *Statistical models and methods for lifetime data*. Wiley, United States.

Nakano, E. Y. (2017). Apostila de análise de sobrevivência. Universidade de Brasília - UnB.

Nobre, L. P. (2016). Modelos de regressão weibull para dados discretos em análise de sobrevivência. Monografia (Graduação em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília.

SABINO, J. L. (2017). Zonas especiais de interesse social - zeis. *Revista Âmbito Jurídico*. Disponível em: http://www.ambito-juridico.com.br/site/?n_link=revista_artigos_leitura&artigo_id=19023&revista_caderno=4.

Santos, D. F. (2017). Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília.

Anexos

A.1 Estimação do modelo

```
#### FUNÇÃO DE MÁXIMA VEROSSIMILHANÇA ####

vero_log_s_cov<- function(parâmetro){
alpha<- parâmetro[1]
gama<-parâmetro[2]
if (gama>0){
p<-(1/(1+(tempo/alpha)^gama))-(1/(1+((tempo+1)/alpha)^gama))
sob<-  1/(1+((tempo+1)/alpha)^gama)
vero<- - sum(censura*log(p)+(1-censura)*log(sob))
}
}

est_log_s_cov<- optim(c(2,1),vero_log_s_cov,hessian = T)
```